About these notes

These are brief notes on the relationship between sample error and true error of a hypothesis, and confidence intervals (based mostly on Mitchell, Chapter 5).

Terminology

We're now working in the classification world. Assume binary classification problems.

Sample error with respect to some sample S is the number of misclassified examples divided by the size of S. If S is the test set, this is the test error.

True error is the probability that a hypothesis will misclassify a single instance drawn at random from the underlying (unknown) distribution D of examples.

We'd like to know the true error, but we can't.

Confidence Intervals

Instead, we try to answer the question of how good an estimate of true error is provided by sample error.

Assuming that examples in S are independently sampled, $|S| = n$, and r errors are made:

$$\sigma_{\text{error}_S(h)} = \frac{\sigma_r}{n} = \sqrt{\frac{p(1-p)}{n}}$$

We don't know p, but we can approximate it by $\text{error}_S(h) = r/n$ (works well for $n > 30$).

Then

$$\sigma_{\text{error}_S(h)} \simeq \sqrt{\frac{\text{error}_S(h)(1-\text{error}_S(h))}{n}}$$

Now, move from the binomial model to a normal approximation (works well for $n > 30$ again. The confidence interval is computed as the mass of the distribution that is within some number z of the mean. For 95% c.i.s $z \simeq 1.96$, giving us the c.i. for the true error:

$$\text{error}_S(h) \pm 1.96\sqrt{\frac{\text{error}_S(h)(1-\text{error}_S(h))}{n}}$$

Question

How many samples do we need to get a good estimate (within a few percentage points) for the probability of a Bernoulli r.v.?